# AN EFFICIENT AND ROBUST HUMAN CLASSIFICATION ALGORITHM

*Yang Ran, Isaac Weiss, Qinfen Zheng and Larry. S Davis*
*Center for Automation Research*
*University of Maryland,*
*College Park, MD 20742-3275, USA*
*{rany, weiss, qinfen, lsd}@cfar.umd.edu*

## ABSTRACT

*This paper describes an object classification algorithm for infrared videos. Given a detected and tracked object, the goal is to analyze the periodic signature of its motion pattern. We propose an efficient and robust solution similar to frequency estimation techniques in speech processing. Periodic reference functions are correlated with the video signal. In order to capture the frequency response at a given set of period, we explore a local version of DFT. By estimating the periodicity at every pixel, we obtain the overall response for the object, which helps us to make decision robustly. Experimental results for both infrared and visible videos acquired by ground-based as well as airborne moving sensors are presented.*

## 1. INTRODUCTION

### 1.1 Motivation

Automatic human activity recognition from video has recently attracted the attention of many researchers [Cutler 2000, Fujiyoshi, 1998, Hogg 1983, 2003]. It plays a critical role in surveillance systems that aim to know what the objects are, and what they are doing [Haritaoglu, 2000]. The periodic nature of human motion has been widely used in gait recognition and related applications [Hogg, 1983; Li, 2002]. The goal of this work is to classify an object as either a human or a vehicle based on its motion pattern.

### 1.2 Related Work

Among the many moving object classification methods, motion signature analysis is a simple and promising approach, especially for infrared and airborne video processing, which typically have low image contrast and small object size. Periodic motion signatures are robust low level clues in these situations.

Frequency estimation in noise-contaminated signal is a well-known problem in signal processing [Rife, 1974]. It is well studied for speech recognition under Gaussian noise assumption. The optimal maximum likelihood estimator (MLE) is obtained by locating the peak in the periodogram. The estimator achieves the Cramer-Rao lower bound for high SNR [Key, 1989]. However, the computational cost is high even with FFT. Additionally, there exists a bias when the signal is not an ideal sinusoid.

Several solutions have been proposed for measuring the periodicity of human motion. [Seitz, 1997] presented a 3-D based detection scheme in curvature space. Polana and Nelson [Polana, 1999] present an approach using DFT. [Efros, 2003] identified the cyclic motion in optical flow domain.

A method closely related to this paper can be found in [1]. The authors used pixel-level correlation to calculate the similarity matrix. Every entry in the matrix represents the similarity between two images of the same object. The periodic property appears as darker lines parallel to the diagonal line (e.g. in Figure 4) and is detected using Short Time Frequency Analysis [Cutler, 2000]. Another commonly used method is to use segmented silhouettes. [Fujiyoshi, 1998] provided a real-time method based on image skeletonization. It uses a 'star' model extracted from a detected silhouette to describe the targets' contour distribution. The evolution of "star" over time reveals the underlying human body motion.

There are limitations to the above approaches. The approach in [Cutler, 2000] requires calculation of a similarity matrix between all pairs of images, which is computationally expensive. Another problem is that it is sensitive to object misalignment as well as changing background. The skeletonization method relies on contour extraction and is sensitive to the quality of silhouette generated. Silhouette detection is a challenging task especially when video contrast is low,

# Report Documentation Page

*Form Approved*
*OMB No. 0704-0188*



| 1. REPORT DATE | 2. REPORT TYPE | 3. DATES COVERED |
|---|---|---|
| **00 DEC 2004** | **N/A** | **-** |

| 4. TITLE AND SUBTITLE | 5a. CONTRACT NUMBER |
|---|---|
| **An Efficient And Robust Human Classification Algorithm** | 5b. GRANT NUMBER |
| | 5c. PROGRAM ELEMENT NUMBER |
| 6. AUTHOR(S) | 5d. PROJECT NUMBER |
| | 5e. TASK NUMBER |
| | 5f. WORK UNIT NUMBER |
| 7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Center for Automation Research University of Maryland, College Park, MD 20742-3275, USA** | 8. PERFORMING ORGANIZATION REPORT NUMBER |
| 9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) | 10. SPONSOR/MONITOR'S ACRONYM(S) |
| | 11. SPONSOR/MONITOR'S REPORT NUMBER(S) |

12. DISTRIBUTION/AVAILABILITY STATEMENT
**Approved for public release, distribution unlimited**

13. SUPPLEMENTARY NOTES
**See also ADM001736, Proceedings for the Army Science Conference (24th) Held on 29 November - 2 December 2005 in Orlando, Florida. , The original document contains color images.**

14. ABSTRACT

15. SUBJECT TERMS

| 16. SECURITY CLASSIFICATION OF: | | | 17. LIMITATION OF ABSTRACT | 18. NUMBER OF PAGES | 19a. NAME OF RESPONSIBLE PERSON |
|---|---|---|---|---|---|
| a. REPORT | b. ABSTRACT | c. THIS PAGE | | | |
| **unclassified** | **unclassified** | **unclassified** | **UU** | **7** | |

the object is small (for example $10\times10$ pixels as in the applications addressed here), or the camera is moving.

Our goal is to develop a computationally efficient and robust periodicity motion analysis algorithm which works well in infrared and airborne videos. In such situations, it is very hard to obtain an accurate silhouette of moving objects. A periodic signal will have peaks at multiples of the period in its power spectral density (PSD). We compare *a set of a priori* signals having some specific periods with the original wave formed by the evolution of a pixel along the temporal axis. When they match, the cross-correlation will approach the maximum. The reference signals are designed based on pattern of typical human motion.

### 1.3 Assumptions

We assume that the moving objects have been detected and tracked. They are specified by bounding boxes in each frame. We assume that the detected objects have been normalized to a fixed size. No constraints on the background are made.

### 1.4 Brief Algorithm Overview

The main idea includes using a finite frequency set to probe the images of an object for its periodic signature and using the period and its strength for classification. A concise signal is derived from the periodic and symmetrical nature of human motion as an *a priori* reference. The method is *efficient* due to low computation cost. The period detection is transformed into a global-maximum location process. It works well for low contrast and small size targets where other methods have difficulties.

## 2. PERIODICITY ANALYSIS BY FINITE FREQUENCIES PROBING

### 2.1 Periodic Signal Probing

The idea of using a reference signal to correlate with the original target signal derives from Fourier domain analysis. An illustration is in Fig. 1. Shown in Fig. 1 (A) is a signal consisting of the superposition of two cosines and the magnitude of its power spectrum (DFT coefficients). There are two peaks corresponding to the two base frequencies.

In the ideal case, the power spectrum of a periodic signal shows peaks at multiples of that period. In reality, we only have a finite length of signal. And due to noise, the signal will be corrupted and the period in the frequency domain will be difficult to detect as shown in Fig 1 (B). The repeated peaks quickly attenuate.

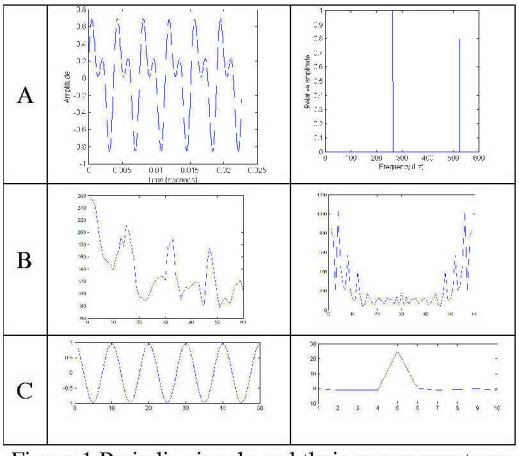

Figure 1 Periodic signals and their power spectrum

The basic idea is to focus on responses at a finite set of frequencies, which are likely associated with the period of human motion. Let us first explore a simple example. Let $W(t)$ be the signal $W(t)=\cos(\omega t)$. Its discrete form is $W(n)=\cos(\omega * n)$. We use another cosine signal with frequency $\omega'$ to correlate with $W(n)$:

$$C(W,W')=\sum_{n=1}^{N}\cos(\omega n)\cos(\omega' n) \qquad (1)$$

where N is the signal length. Row C of Fig. 1 gives the response at different ? '. It is clear that the response will have a *global* maximum at $\omega$, even if we only use a signal with finite number of periods. The probing is defined as follow.

***Definition 1**: Probing is a process of matching a periodic reference signal to the target signal to obtain a measure of their correlation.*

Starting with a quasi-periodic signal $W(t)\approx W(t+nT)$, where T is the quasi-period of the signal. If we use a temporal window to truncate the sample of $W(t)$, we will get a vector $\overline{W}(t)=[W(t),W(t+\tau)...W(t+(N-1)\tau)]$. Given a reference signal $W'$ and under additive Gaussian noise, at T the following *a posteriori* probability is maximized

$$P_{w|H_T}(W(t)\mid H_T)=\frac{1}{(2\pi\sigma^2)^{N/2}}\left(-\frac{\|W(t)-s(t,T)W'(t)\|^2}{2\sigma^2}\right) \qquad (2)$$

where $s(t,T)$ is a scaling function and $H_T$ is the hypothesis that period is T. Partial differentiation gives

$$\frac{\partial}{\partial s(t,T)}\left(-\frac{\| W(t)-s(t,T)W'(t)\|^2}{2\sigma^2}\right)=0 \Rightarrow s(t,T)=\frac{W(t)W'(t)}{\|W(t)\|\|W'(t)\|} \quad (2')$$

Using Bayes rule we have

$$P_{H_T|W}(H_T \mid W(t)) = \frac{P_{s(t)|H_T} P_{H_T}}{P_{W(t)}} \quad (3)$$

The best estimate of the quasi-period is the frequency which maximizes the cross correlation $C(W,W')$.

### 2.2 Probing with Finite Frequency Sets

There are two major concerns for real signals. One is segmentation of object from the background. Infrared videos usually have low contrast. In some cases the camera is moving, so we cannot use the background subtraction method [Culter, 2000]. The other concern is computation cost.

The output of the detecting/tracking module gives a sequence of bounding boxes for every object. After alignment, we will have a stack of rectangles with the same size and object center. A probing function with period ? and waveform $k$ is defined as $\Phi_k(\omega)=k(\omega t)$. The overall cost function is defined over the whole definition space of $W(t,x,y)$, where $W(t,x,y)$ is the pixel or a corresponding feature at location $(x,y)$ at time t.

$$C(k,\omega)=\int_x\int_y\int_{-\infty}^{\infty}\Phi_k(\omega)\cdot W(t,x,y)dtdydx \quad (4)$$

In practice, we have limited length and size and the discrete version is:

$$C(k,\omega)=\sum_{x=1}^{X}\sum_{y=1}^{Y} cor(\Phi_k(\omega),W(n,x,y)) \quad (4')$$

Our goal is to calculate the overall correlation of the signal W by summing up the value at each location (x, y) with a same reference signal F at frequency ?. The period is defined from (3) and (4') as the ? in $(\omega_1,\omega_2,\ldots,\omega_m)$, which maximizes the averaged response function (4).

$$period=\arg\max P_{H_T|W}(H_T \mid W(t)) = \arg\max_{\varpi\in\{\varpi 1,\varpi 2..\varpi m\}} C(k,\omega)$$
$$= \arg\max_{\varpi\in\{\varpi 1,\varpi 2..\varpi m\}} \sum_{x=1}^{X}\sum_{y=1}^{Y} cor(\Phi_k(\omega),W(n,x,y)) \quad (5)$$

### 2.3 Periodicity Detection

To detect the period efficiently, we need to select the appropriate probing function k. Given a signal, the ideal probing function is the signal itself, because the correlation will approach maximum. Since the input signal is not available in advance, several possible functions are tested and compared. By intuition, the triangle and cosine/sine functions appear to be appealing due to their simple and representative form. Analysis of typical human walking will give more suitable reference signals.

We have observed that a video of walking pedestrian will have both period and symmetry due to movement of legs and arms. This can be used to design the reference signal. Observations from [Cutler, 2000, 1998] suggest designing a twin-peak reference signal representing both periodicity and symmetry. Figure 2 illustrates these two properties. It is a complete cycle of a walking human. In addition to the similarity between cycles, there is also resemblance between the first and the second halves, shown as two rows in Figure 2.

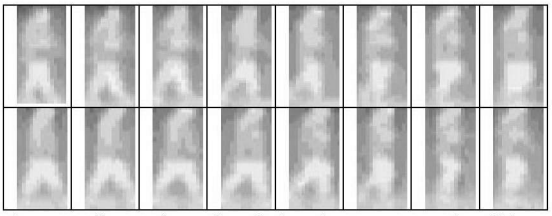

Figure 2 Illustration of period and symmetry of walking

We investigated the similarity between pedestrian sequences in left two images of Figure 3 using the method in [Cutler, 2000]. After simple smoothing, we noticed that there are two peaks in every cycle due to period and symmetry mentioned above.

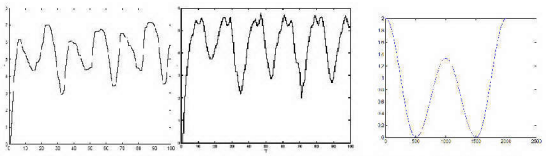

Figure 3 Similarity signals and twin-peak signal

Hence we use a twin-peak reference signal for probing. It is generated by combining two sinusoids as shown in the right image in Figure 3. The first peak, due to period, appears at the multiples of period T and the second, due to symmetry, at $(n+1/2)T$.

The probing results for pedestrians are shown in Figure 4. Two examples are presented. One is a ground-based infrared video and the other is an airborne video. The second row of Figure 4 shows the correlation matrices calculated as in [Cutler, 2000]. Although there are darker lines parallel to the diagonal line in the airborne data

corresponding to the period, we cannot find any in the correlation matrix for a walking human in infrared video. Yet the probing method detects distinct peak at the periodic frequency as shown in the third row in Figure 4.

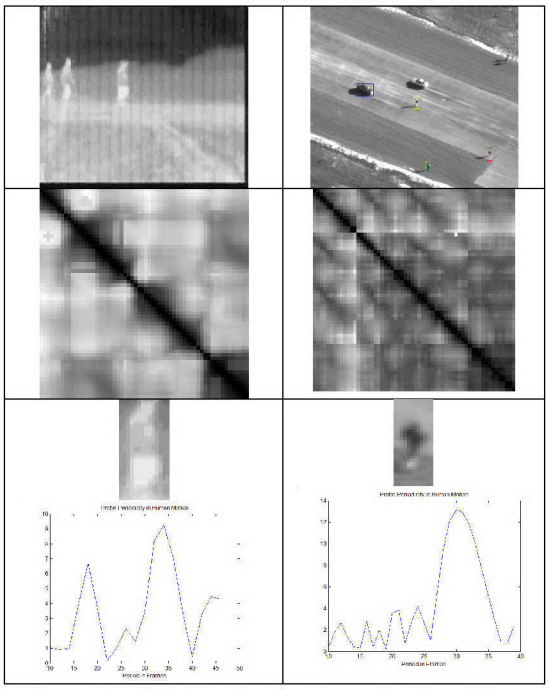

Figure 4 Period detection for two pedestrian videos.

In order to have a better idea about how well this method works, we plot the intensity change in Figure 5 for a column of pixels (They are in the dark line in the right image) from a human in left column of Figure 4. Although the period is only in several pixels and is even hard to identify by eyes, the proposed algorithm successfully detects it.

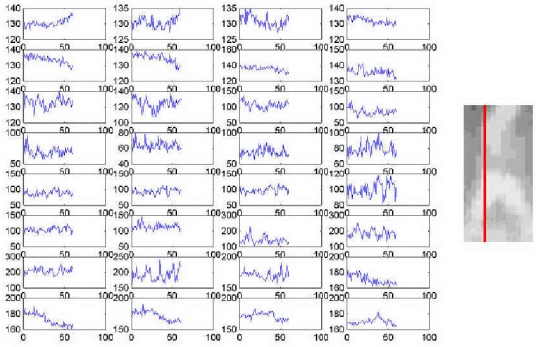

Figure 5 Intensity evolution for pixels along a column

### 2.4 Object Classification

Period detection now becomes a simple process as finding a global maximum as defined in (5) within the finite frequency set, which could be implemented with many fast and reliable methods. The decision as to whether or not a moving object is a person is made as follows: a candidate object is classified as a human if and only if at some time $t$, the global peak satisfies

$$\frac{Peak - Mean}{Variance} > TH \qquad (7)$$

In Equation (7), *Peak* is the maximum value over all probing frequencies. *Mean* and *Variance* is the standard statistics of all the correlation coefficients excluded those located within a local support window with a given size centered at the index giving maximum. *TH* is the threshold for a confident decision. In our experiments, typical values are between 2 to 4.

## 3. EXPERIMENTAL RESULTS

This section describes the classification results of the proposed algorithm. It includes details on the structure of the data as well as results on both infrared and visible videos.

### 3.1 Structure of Testing Data

Two datasets were tested. The first dataset is from infrared ground sensors. This set consists of 10 sequences containing more than 20 clips (15 pedestrians and 5 vehicles). The foreground objects include typical scenes such as a parking lot, a road, and other urban scenarios. There are different objects with various sizes and poses. The second dataset is gray level airborne videos. There are 10 human sequences and 5 vehicle sequences. All data are captured at a speed of 30 frames per second. For both datasets, the detection and tracking algorithms we used are reported in [Culter, 2000; Haritaoglu, 2000].

### 3.2 Experiments on Infrared Videos

In Figure 6, we present the probing process at a given time. The video length is 60 frames, which corresponds to about two cycles of a normal walking pedestrian. The result is shown for detecting the periodic motion for different targets. The blue line is the plotted response and the red line is the mean of the response over all frequencies. These are humans walking through an urban scene at different poses. All of them have a dominant twin-peak correctly classified using (7).

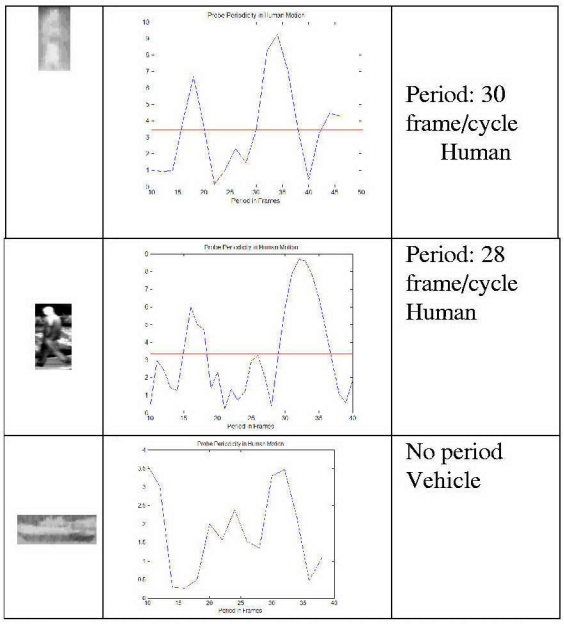


Figure 6(a) Probing for different objects in infrared videos.

Then we perform the similar algorithm on ground based video and present the result in Figure 6(b).

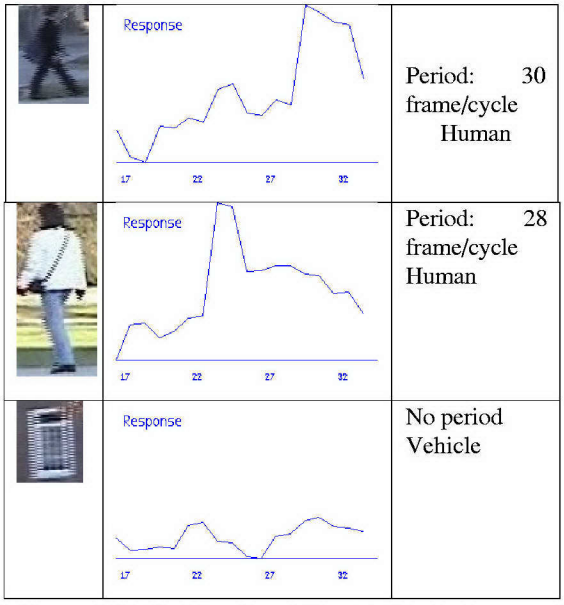


Figure 6(b) Probing for different objects in ground based videos

For more reliable classification, we may check the consistency of the periodicity over time. In Figure 7, a continuous human period detection over a long time interval is shown. The left image is the superimposed response for 5 different times for a 100-frame pedestrian sequence. Not only is the twin-peak pattern distinct, but also the period consistently falls into a narrow range around 25 frames/cycle, which is shown in the right image. By checking this consistency, the object can be classified with high confidence as a pedestrian.

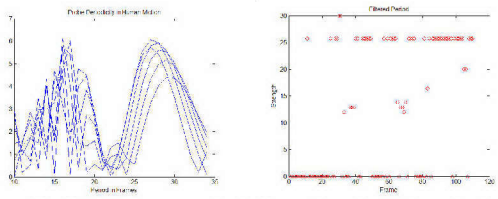

Figure 7. Continuous period detection in infrared data. (Left) responses over five different time periods; (Right) plot of detected periods at different time.

### 3.3 Experiments on Airborne Videos

In the airborne sequence, targets are moving humans and vehicles on the ground. Object sizes are less than 10x10 pixels. Leg periodic motions are less obvious in overhead views. Figure 8 and Figure 9 show experiment results. The proposed method manages to classify the targets correctly.

Figure 8 provides classification results for two human and a car. In Figure 9, the continuous probing result is shown for a human and a vehicle. Unlike the case for human, the detected periods for a vehicle are randomly distributed in a wide range.

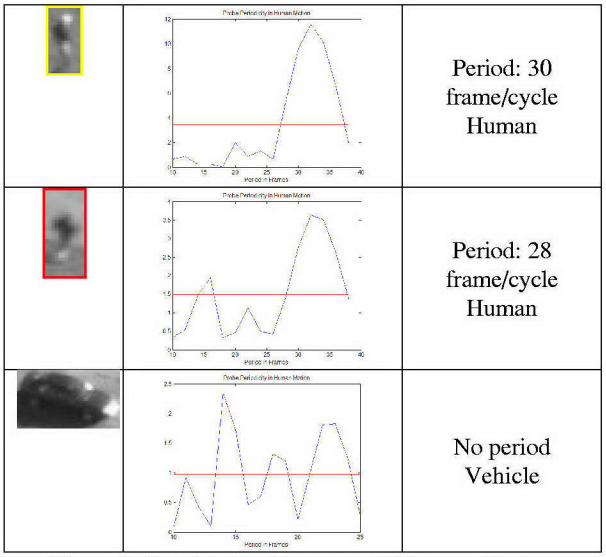


Figure 8 Results for airborne surveillance video

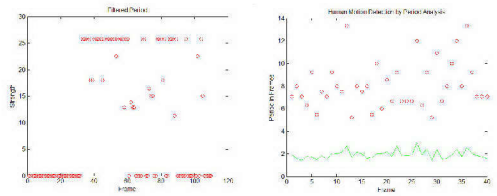

Figure 9 Continuous period detection for human (left) and a car (right) in airborne videos.

## 4. SENSITIVITY ANALYSIS

We have tested several key factors of the algorithm which may affect the result. In the evaluation, we pay special attentions to following two questions:

- Will still detect the twin pattern when we change a key experimental factors;
- To what extend will the result be affected in terms of peak/mean ratio.

We use a variable $C$ defined as the change of the peak/mean ratio in percentage when we change a condition.

$$C = abs(P/M - P'/M')/(P/M) \times 100\% \quad (8)$$

where P, M is the original peak and mean value and P', M' is the new values after changing some factors.

### 4.1 Computational Cost

The first factor we are concerned with is the speed of the method. Suppose the bounding boxes after alignment have a width w and a height h, and the video length is N frames. The probing frequency ranges over n frequencies. Then the required addition and multiplication operations are N*w*h*n respectively. The time is given by:

$$T = N \cdot w \cdot h(ADD + MUL) \cdot n$$

which is a linear computation time in terms of N or n. The processing time can be further reduced by using multi-resolution probing. A coarse frequencies set is first used to roughly locate the global maximum. Then a denser set is generated around the detected frequency and used to obtain refined frequency.

### 4.2 Alignment

The cost function requires good alignment of the frames for the detected objects since it use pixel wise temporal correlation. Current detecting and tracking algorithms cannot provide error-free alignment for bounding boxes. To get a quantitative comparison, we add Gaussian noise to a set of calibrated bounding boxes. By increasing the variance, we measure the peak/mean ratio change compared to the original result. Table 1 shows the result.

Table 1. Comparison at different alignment error

| S | 0.5 | 1.0 | 1.5 | 2.0 | 2.5 | 3.0 |
|---|---|---|---|---|---|---|
| Period | 34 | 34 | 33 | 34 | 32 | 36 |
| C(%) | 95.6 | 78.9 | 50.2 | 37.1 | 12.2 | 5.8 |

### 4.3 Object Size

We give the measurement for one sequence with different down-sampled sizes. This method exhibits a strong invariance to the object size. We obtain good results even when the target size is reduced to 10x10. Notice that during the sub-sampling, the detected period does not change. We give the change in peak/mean ratio in Table 2.

Table 2. Comparison at different target sizes (original is 100x80)

| Sub-sample ratio | 2 | 4 | 6 | 8 |
|---|---|---|---|---|
| C(%) | 93.2 | 87.9 | 76.3 | 70.1 |

### 4.4 Video Length

An interesting issue is the minimal length needed to analyze the period with sufficient confidence. This is equivalent to considering the size of window we use to truncate the signal. Suppose we estimate the frequency directly from FFT result without any further processing [Kay, 1989]. If the true period is ? , and it falls into two adjacent bins: k and k+1,

$$\omega \in [k * 2\pi F_{sample}/N, (k+1) * 2\pi F_{sample}/N] \quad (9)$$

where $F_{sample}$ is the sample frequency, we will have a bias up to the width of the bin. Hence this method requires longer sequence for higher resolution. But this is determined by the tracking algorithm, and thus it is not always easy to achieve in low quality video, small object, and from a moving platform. In the proposed method, for a typical human, we need about two to three cycles (60-90 frames for a 30 fps video) to estimate the correct period.

Besides, the cross-correlation of the two signals will attain maximum when the windows (length) is a multiple of the period. Hence there would be residue in (5) if this is not true. The error can be suppressed by summation over all pixels, which is a topic for future work.

### 4.5 Frame Rate

Due to sensor limitation, we may be unable to get the full frame rate all the time. In addition, robustness to frame rate drop could be useful for saving overall computational cost.

Table 3. Comparison at different frame rates

| frame rate | 20 | 15 | 10 | 5 |
|---|---|---|---|---|
| period | 17(34) | 16(32) | 11(33) | 5(25) |
| C(%) | 95.0 | 87.9 | 46.3 | 20.1 |

In the second row, the number in parenthesis is the normalized period of the original frame rate. The results show that the probing is more sensitive to down sampling in object size than in frame rate.

## 5. SUMMARY AND DISCUSSION

A periodicity motion detection based object classification algorithm is reported. The method is simple, efficient, and robust to target size and frame rate. It transforms the complicated period detection into an easier global maximum location process. The choice of the probing by *a priori* reference signal within a finite frequency set enables accurate object classification even with a short video clip (2-3 seconds). Sensitivity analysis reveals that robust nature of the proposed method.

## 6. REFERENCES


Cutler, R. and Davis, L.S. 2000, "Robust real-time periodic motion detection, analysis, and applications," *IEEE Trans. PAMI*, **22**(8):781-796.

Cutler, R. and Davis, L.S., 1998, "View-based detection and analysis of periodic motion," *Proc. ICPR*, **1**:495-500.

Efros, A.A., Berg, A.C., Mori, G. and Malik, J., 2003, "Recognizing action at a distance," *Proc. ICCV*, pp. 726-733.

Fujiyoshi, H. and Lipton, A.J.. 1998, "Real-time human motion analysis by image skeletonization," *Applications of Computer Vision, Proc.* WACV pp.15–21.

Hogg, D. 1983, "Model-based vision: A program to see a walking person," *Image and Vision Computing*, **1**:5-20.

Hogg, D. Binefa, X., 2003, "Classifying periodic motions in video sequences," *Proc. ICIP*, **1**:945-948.

Haritaoglu, I., Harwood, D., and Davis, L.S., 2000, "W4: Real-time surveillance of people and their activities," *IEEE Trans. PAMI*, **22**(8):809-830.

Kay, S., 1989, "A fast and accurate single frequency estimator," *IEEE Trans. SP*, **37**(12): 1989.

Li, B. Holstein, H., 2002, "Recognition of human periodic motion-a frequency domain approach," *Proc. ICPR*, **1**:311-314.

Polana, R. and Nelson, R.C. 1992, "Recognition of motion from temporal texture," *Proc. CVPR,* pp.129-134, Champaign, Illinois.

Rife, D.C., and Boorstyn, R.R., 1974, "Single-tone parameter estimation from discrete-time observations," *IEEE Trans. IT*, **20**(5):591-598, Sep.

Seitz, S.M., Dyer, C.R., 1997, "View-invariant analysis of cyclic motion," *Intl Journal of Computer Vision*, 25:1-23.